In: Proc. ESCA Workshop Auditory Basis Speech Perception, pp. 317-320, 1996

CHANNEL NORMALISATION BY USING RASTA FILTERING AND THE DYNAMIC CEPSTRUM FOR AUTOMATIC SPEECH RECOGNITION OVER THE PHONE

Boda Peter-Pal(1), Johan de Veth(2) & Louis Boves(2,3)
(1) Laboratory of Acoustics and Audio Signal Processing,
Helsinki University of Technology, Otakaari 5 A, SF-02150 Espoo, FINLAND
(2) Department of Language and Speech, University of Nijmegen,
P.O. Box 9103, 6500 HD Nijmegen, THE NETHERLANDS
(3) KPN Research, P.O. Box 421, 2260 AK Leidschendam, THE NETHERLANDS

ABSTRACT

Human auditory perception is perfectly capable to deal with time-invariant linear filter effects, such as those introduced by telephone handsets and telephone channels. We compared two different schemes for modeling human auditory time-frequency masking: RASTA filtering and the dynamic cepstrum representation (DCR). We used a small set of context-independent phone hidden Markov models for a recognition task of connected digit strings over the telephone. We found that RASTA filtering outperformed the Gaussian DCR approach, despite the fact that RASTA represents a more crude approximation of human forward masking. Our results may be influenced by the choice of the mel-frequency cepstral representation that we used. The superiour performance of the RASTA technique may also be explained by the fact that the frequency response of the RASTA filter is better matched to the region of modulation frequencies where human auditory perception is most sensitive.

1. INTRODUCTION

For automatic speech recognition over the phone, the filtering effects of the handset microphone and the telephone channel may cause serious performance degradation. However, recognition by humans is very robust against stationary distortions of the speech signal. It has been suggested that the well documented time-frequency masking properties of the human auditory system contribute to this robustness. In this paper we discuss how models of human auditory time-frequency masking can be used in the acoustic front-end of an automatic speech recognizer to enhance channel robustness. We compare two different schemes for modelling time-frequency masking in the context of a connected digit recognition task: RASTA filtering [1] and the dynamic cepstrum representation [2].

The linear filtering characteristics introduced by the communication channel can be modelled in the cepstral domain as a constant additive bias vector. As has been suggested in [1], RASTA filtering a sequence of cepstral vectors effectively removes the DC-component, that can be attributed to the linear filtering characteristics introduced by the the communication channel and the overall shape of the vocal tract of the speaker. It has been repeatedly shown that RASTA filtering substantially improves recognition performance when word models or context dependent models are used [1, 3, 4]. However, RASTA may be less suited for context independent (CI) modeling due to the strong left context dependency introduced by the long filter memory. The most important effect of the RASTA filter is its high-pass characteristic at low modulation frequencies. As argumented recently [5], RASTA filtering may be interpreted as a crude means to approximate the temporal (forward) masking effect in human auditory perception.

A somewhat more sophisticated approach to model temporal masking in human auditory perception has been proposed by [2]. These authors introduced the dynamic cepstrum representation (DCR) as a model of forward auditory time-frequency masking and showed that consistently better recognition performance was achieved when comparing the dynamic cepstrum to the conventional cepstral representation. The ability of the dynamic cepstrum to enhance the spectral dynamics and to reduce the effects of the overall average spectrum make this approach also an attractive candidate for channel normalisation.

In this paper we describe a number of experiments where we trained context-independent (CI) phone model HMMs. We compared the recognition performance in case RASTA filtered and Gaussian DCR acoustic vectors were used. This paper is organised as follows. In sections 2 and 3 the signal processing technique and the speech material are described. In section 4, we discuss the type of HMMs, the cross-validation scheme to train these and the recognition syntax for our experiments. In section 5, we discuss the results for the different channel normalisation techniques that we studied. Finally, the conclusions are presented in section 6.

2. SIGNAL PROCESSING

Speech signals were digitized at 8 kHz and stored in A-law format. After conversion to a linear scale, preemphasis with factor 0.98 was applied. A 25 ms Hamming analysis window that was shifted with 10 ms steps was used to calculate 24 filterband energy values for each frame. The 24 triangular shaped filters were uniformly distributed on a mel-frequency scale. Finally, 12 mel-frequency cepstral coefficients (MFCC's) were derived. We did not apply liftering, because we were using continuous Gaussian mixture density HMMs with diagonal covariance matrices. For this class of models, multiplying a coordinate of the feature vector with some constant (say c) is equivalent to multiplying the probability used during the dynamic programming (DP) with c^{-1}. In other words: the factor $\log(c)$ is added to each negative log probability in the DP. Since this factor occurs in each negative log prob, the result of the DP is not affected. In addition to the twelve MFCC's we also used their first time-derivatives (delta-MFCC's), log-energy (logE) and its first time-derivative (delta-logE). In this manner we obtained 26-dimensional feature vectors. Feature extraction was done using HTK v1.4 [6].

For channel normalisation we applied the filtering (either the RASTA filter with integration factor 0.98 [1] or the Gaussian

digit	transcription	trn960	trn480	tst911	tst240
nul	n Y l	590	294	548	136
een	e n	590	286	562	165
twee	t w e	591	296	597	181
drie	d r i	597	299	574	155
vier	v i r	569	284	523	135
vijf	v Ei f	573	273	526	124
zes	z E s	578	301	536	136
zeven	z e v Q n	582	270	510	130
acht	a x t	554	297	525	151
negen	n e x Q n	534	281	556	121

Table 1: Phonemic transcriptions (column 2) and the number of realisations (columns 3 till 7) of each digit.

DCR filters as defined in [2]) to the twelve MFCC feature coordinates only. If it can indeed be safely assumed that the effect of a channel is an additive constant in the cepstral domain [1], then the delta-MFCC coefficients are already robust with respect to the type of channel mismatches that RASTA and DCR can compensate for. Also, we kept the original values of logE and delta-logE.

3. DATABASE

The speech material for this experiment was taken from the Dutch POLYPHONE corpus [7]. Speakers were recorded over the public switched telephone network in the Netherlands. Handset and channel characteristics are not known; especially handset characteristics are known to vary widely. The speakers were selected in such a way that all major dialect backgrounds in the Netherlands are represented. None of the utterances used for training or test had a high background noise level.

Among other things, the speakers were asked to read a connected digit string containing six digits. We divided this set of digit strings in two parts. For training we reserved a set of 960 strings, i.e. 80 speakers (40 females and 40 males) from each of the 12 provinces in the Netherlands (denoted trn960 in short). An independent set of 911 utterances (tst911; 461 females, 450 males) was set apart for testing. (In principle we again wanted to have 40 female and 40 male speakers from each of the 12 provinces, but the very sparsely populated province of Fleveoland provided only 21 female and 10 male test speakers). For proper initialisation of the models, we manually corrected automatically generated begin- and endpoints of each utterance in the trn960 data set.

We did not always use all training and testing material. Most of the time, we used only half the amount of training data (i.e. 480 utterances, trn480; 240 females, 240 males). For cross-validation during training we used a subset of 240 utterances taken from the test set (tst240; 120 females, 120 males). For evaluation of the models when training was completed we always used the full test set tst911. We list the number of available realisations of each digit for all of our data sets in columns 3 till 6 of Table 1.

4. MODELS

4.1. Model topology

The digit set of the Dutch language was described using 18 context independent (CI) phone models (see second column of Table 1). Furthermore, we used four models to describe silence, very soft background noise, other background noise and out-of-vocabulary speech, respectively. Each CI model consists of a three state, left-to-right HMM, where only self-loops and transitions to the next state are allowed. The emission probability density functions are described as a continuous mixture of 26-dimensional Gaussian probability density functions (diagonal covariance matrices). In order to be able to study the recognition performance as a function of acoustic resolution, we used mixtures containing 1, 2, 4, 8, 16 and 32 Gaussians for the emission probability density function of each state. In this manner, the total number of Gaussians ranged from 66 (in case of the single Gaussian models) to 2112 for the models with 32 Gaussians per state.

4.2. Training

The CI phone models were initialised starting from a linear segmentation within the boundaries taken from the hand-validated word segmentations. After this initialisation, an embedded Baum-Welch re-estimation was used to further train the models. Starting with a single Gaussian emission probability density function for each state, 20 Baum-Welch iterations were conducted; the models resulting from each iteration cycle are stored. Next, the optimal number of iterations was determined using the tst240 data set. For the set of models with the best recognition rate, the number of Gaussians was doubled and again 20 embedded Baum-Welch re-estimation iterations were performed. This process of training with cross-validation was repeated until models with 32 Gaussians per state were obtained.

4.3. Recognition

During cross-validation as well as during recognition with data set tst911, the recognition syntax allowed for zero or more occurrences of either silence or very soft background noise or other background noise or out-of-vocabulary speech in between each pair of digits. At the beginning and at the end of the digit string one or more occurrences of either silence or very soft background noise of other background noise or out-of-vocabulary speech were allowed.

5. EXPERIMENTS

5.1. No channel normalisation

In order to establish a reference for the connected digit recognition task that we used, we trained models with up to 32 Gaussians per state using data set trn480. The feature vectors were not channel normalised in any way. The recognition results for the cross-validation data set tst240 are shown in Figure 1. From this figure it can be deduced, that doubling the number of Gaussians per state from 1 to 2 was performed using the models obtained at iteration 20, doubling from 2 to 4 at iteration 20, 4 to 8 at 4, 8 to 16 at 12 and 16 to 32 at 20. Our data did not show any system with respect to the number of iterations that yielded the best performing models. What was systematic, however, was the occurrence of many cases where cross-validation recognition rate showed local maxima comparable to the one at 7 iterations in the bottom curve in Figure 1.

The best performing model sets were evaluated using test set tst911. The proportion of digits correct (i.e. the number of digits correctly recognized divided by the total number of digits in the test set) is shown as a function of the total number of Gaus-

sians in all HMMs in Figure 2. In this figure the 95% confidence intervals are indicated as vertical bars.

From Figure 2, it appears that 528 Gaussians (i.e. 8 Gaussians per state) is the optimal acoustic resolution required for the given amount of training data that we used. Although the recognition performance increases monotonically as a function of the acoustic resolution, the improvements are not significant for 16 and 32 Gaussians per state. We found a word error rate (WER; defined as one minus the proportion of digits correct) of $3.7 \pm 0.5\%$ at 8 Gaussians per state.

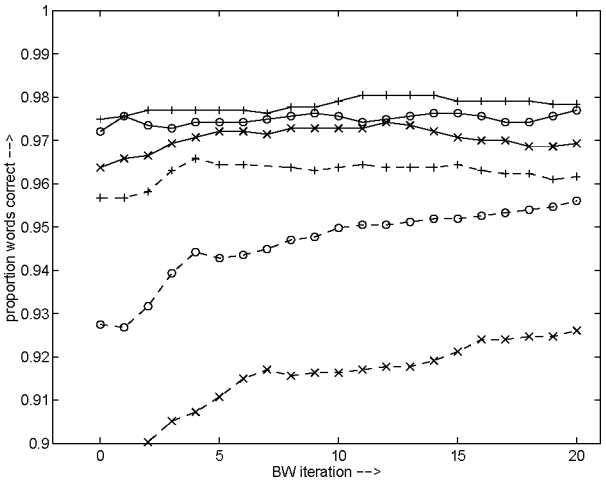

Figure 1: Proportion of words correct as a function of the number of Baum-Welch iterations. Dashed lines: $\times, O, + = 1, 2, 4$ Gaussians per state. Solid lines: $\times, O, + = 8, 16, 32$ Gaussians per state.

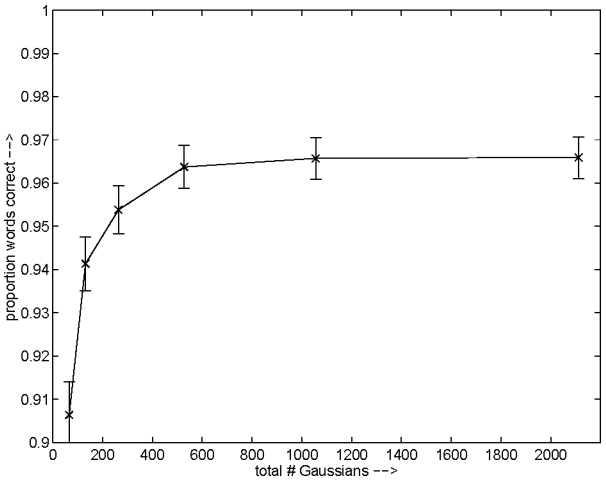

Figure 2: Recognition performance for test set tst911 as a function of the acoustic resolution of the models.

5.2. Channel normalisation

We replaced the twelve cepstral coefficients in our feature vectors by twelve RASTA filtered cepstral coefficients. In a separate experiment we calculated the Gaussian DCR for each of our twelve cepstral coefficients. In both cases, we retrained HMMs with the new feature vectors using exactly the same utterances for training and cross-validation. Again, the best models sets according to test set tst240 were avaluated with test set tst911. The recognition results are shown in Figure 3 together with the reference results. Figure 3 clearly indicates that both channel normalisation techniques improve the recognition performance: the proportion of digits correct is larger for each number of Gaussians that we tested. Notice that the improvements by using the RASTA filtering technique are significant at the 95% confidence level, but that those of the Gaussian DCR approach are not. Furthermore, the difference between RASTA and Gaussian DCR is significant. At 8 Gaussians per state we found WER $= 3.3 \pm 0.5\%$ for Gaussian DCR features and WER $= 2.6 \pm 0.5\%$ in case RASTA filtering was applied. In other words, due to the RASTA filtering the WER was reduced by 29% relative to the baseline performance obtained without channel normalisation. A last experiment was performed to verify that we used enough training data. To this aim models were trained with the trn960 data set using the RASTA filtered acoustic vectors. At 8 Gaussians per state, we found WER $= 2.8 \pm 0.5\%$. Because we did not observe a significant change in recognition performance, we concluded that data set trn480 was indeed large enough.

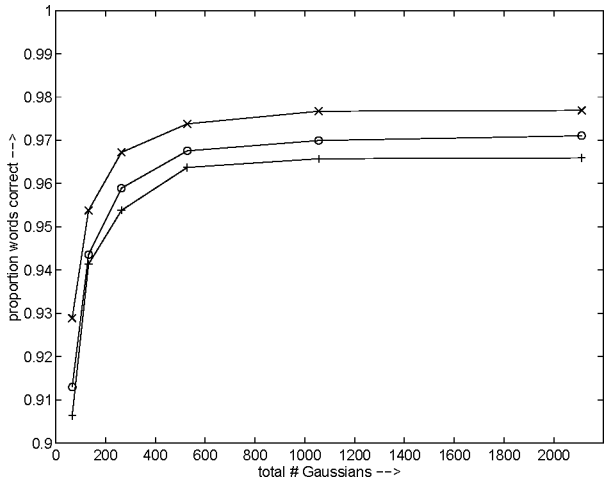

Figure 3: Recognition performance for three types of acoustic features: $\times$ = RASTA filtering, O = Gaussian DCR, $+$ = no channel normalisation.

It has been reported that RASTA filtering did not appear to help increasing performance when using CI HMMs, due to the left-context dependency introduced by the long memory of the RASTA filter (for example see [8]). The fact that we did observe an improvement in our experiments may be explained by the fact that the number of different contexts for each phone model is not very large in our connected digit recognition task. In fact, as can be seen in the second column of Table 1, in 9 cases the phone models occur in one single context only, with an average number of different left contexts for each phone model as low as 1.6.

Because RASTA is a much more crude way to describe human auditory masking, we did not expect to find that RASTA outperforms the Gaussian DCR approach. Maybe the difference was caused by the fact that we used a mel-frequency scale before calculating the cepstral coefficients. In the original paper [2] a linear frequency scale was used. Note that the mel-frequency scale already takes account of frequency masking with zero time lag. So maybe in our Gaussian DCR feature vectors the masking effects may have been over-emphasised, which could lead to performance deterioration.

Our results can also be interpreted as follows. Recently, it has

been shown that human auditory perception is most sensitive in the region of 2-16 Hz [9]. In Figure 4 the frequency response of the RASTA filter and the Gaussian DCR are shown. First, it is important to notice that the attenuation of the DC-component by the RASTA filter is much better than that of the Gaussian DCR. Second, the passband (attenuation less than 3 dB) starts well below 2 Hz for the RASTA technique. However, in case of the Gaussian DCR, the passband starts at 5 to 6 Hz (depending on the cepstral coefficient). This means that especially at low modulation frequencies, the RASTA filter is much better tuned to the maximally sensitive region of human auditory perception. This may explain why we found that RASTA filtering outperformed the Gaussian DCR approach.

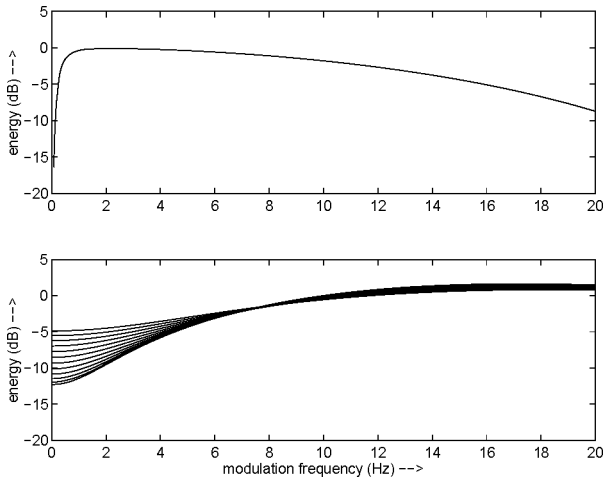

Figure 4: Frequency responses for the RASTA filter (upper panel) and Gaussian DCR (lower panel). In case of the Gaussian DCR the curve starting at -12 dB corresponds to frequency response for the first cepstral coefficient. The frequency response starting at -5 dB corresponds to cepstral coefficient 12.

6. CONCLUSIONS

In this paper we discussed how models of human auditory time-frequency masking can be used in the acoustic front-end of an automatic speech recognizer to enhance channel robustness. We compared two different schemes for modelling time-frequency masking: RASTA filtering and Gaussian DCR. We used a small set of CI phone HMMs for connected digit recognition over the phone. Without an explicit channel normalisation technique, we found WER $= 3.7 \pm 0.5\%$ when using 8 Gaussians per state. Despite the fact that RASTA represents a more crude approximation of human forward masking, RASTA filtering outperformed the Gausian DCR approach: WER $= 2.6 \pm 0.5\%$ vs. WER $= 3.3 \pm 0.5\%$ at 8 Gaussians per state. Our results may be influenced by the choice of the mel-frequency cepstral representation that we used: in its original form, Gaussian DCR modelling was proposed for cepstra based on a linear frequency scale. The superiour performance of the RASTA technique may also be explained by the fact that the frequency response of the RASTA filter is better matched to the region of modulation frequencies where human auditory perception is most sensitive (2-16 Hz).

Acknowledgement

This work was funded by the Netherlands Organisation for Scientific Research (NWO) as part of the NWO Priority programme Language and Speech Technology.

References

[1] H. Hermansky, 'Compensation for the effect of the communication channel in auditory-like analysis of speech', in Proc. Eurospeech-91, Genova, Sept. 1991.

[2] K. Aikawa, H. Singer, H. Kawahara & Y. Tohkura, 'A dynamic cepstrum incorporating time-frequency masking and its application to continuous speech recognition', in Proc. ICASSP-93, pp. 668-671, 1993.

[3] J-C. Junqua, D. Fohr, J-F. Mari, T.H. Applebaum & B.A. Hanson, 'Time derivatives, cepstral normalisation and spectral parameter filtering for continuously spelled names over the telephone', in Proc. Eurospeech-95, pp. 1385-1388, 1995.

[4] H. Singer, K.K. Paliwal, T. Beppu & Y. Sagisaka, 'Effect of RASTA-type processing for speech recognition with speaking-rate mismatches', in Proc. Eurospeech-95, pp. 487-490, 1995.

[5] H. Hermansky & M. Pavel, 'Psychophysics of speech engineering systems', in Proc. ICPhS-95, pp. 3.42-3.49, 1995.

[6] S. Young & P. Woodland, 'HTK v1.4 User Manual', Speech Group, Cambridge University Engineering Department, UK, 1992.

[7] E.A. den Os, T.I. Boogaart, L. Boves & E. Klabbers, 'The Dutch Polyphone corpus', in Proc. Eurospeech-95, pp. 825-828, 1995.

[8] J. Cohen, Final report of the chairman, Frontiers of Speech Processing - Robust Speech Recognition 93, 1993.

[9] R. Drullman, J.M. Festen & R. Plomp, Effect of temporal envelope smearing on speech reception', J. Acoust. Soc. Am., vol. 95, pp. 1053-1064, 1994.